

Cross-Population Comparability: an introduction

Ritu Sadana

Global Program on Evidence for Health Policy, WHO, Geneva

**For presentation during Session 8 of the
Joint UN-ECE and WHO Expert Meeting on Health Status Measurements
Hosted by Statistics Canada
Ottawa, 23-25 October 2000**



**Global Programme on Evidence for Health Policy
World Health Organization
Geneva**

1. Background

The realization that numerous factors contribute to how individuals make judgements and respond within interview based surveys is not new (Jobe & Mingay 1989; Kahneman, Slovic & Tversky 1982). Over the past 50 years or so, investigations in the area of “respondent bias” have documented various sources of error, i.e., factors that may contribute to the difference between what is truth, what is perceived or judged and what is ultimately reported or observed. Earlier work primarily focused on describing these different factors (see Streiner & Norman 1995), while latter work investigated the cognitive processes that contribute to such biases (see Kahneman, Slovic & Tversky 1982). Some of these factors and processes include:

- Social desirability or secondary gains (purposively under- or over-reporting level of health status)
- Aversion to end points or central tendency bias (e.g., reluctance to use the extreme categories on a scale, such as very bad and excellent)
- Halo effect (e.g., the general impression influences how specific traits or states are assessed, and may reflect that raters are unable to evaluate more than a few dimensions distinctly from a general dimension)
- Positive bias (e.g., most likely to agree or acquiesce with interviewer)
- Framing effects (e.g., choice between alternatives depends on how the question is framed rather than the actual content of the question, and the response therefore measures other attributes of the individual such as risk aversion or risk taking, rather than those intended by the original question)
- Judgmental heuristics that are biased (e.g., common strategies people use to make judgements that are influenced by other factors that serve to decrease the accuracy of responses)

Combining perspectives from a variety of fields, including psychology, decision sciences and statistics, a classic article by Tversky & Kahneman (1974) discussed that people rely on a limited number of heuristic principles in order to simplify complex decision making. Three common heuristic principles include *representativeness* (e.g., if asked about A, a person uses information on B, and bases the decision making process concerning A on the degree to which A is representative or resembles B), *availability* (e.g., a person assesses a situation based on the instances or occurrences that can be brought to mind based on one’s own experiences), and *adjustment and anchoring*, (e.g., a person estimates an initial response or anchor starting value, and then during the decision making process makes adjustments – usually partial or incomplete – to that starting value and arrives at a response). Based on many empirical examples, the authors conclude that these heuristic principles are sometimes useful, but can lead to severe and potentially systematic errors.

An optimal strategy for cross-population comparability of data (see Tourangeau 1984, Krosnick 1991, cited in Streiner & Norman 1995) would require that all individuals with the same true level of health status, irrespective of their age, sex, cultural or geographic context, or other socio-demographic characteristics, or time period, respond to an identical question addressing health status as follows:

- interpret the meaning of the question and response scale identically
- retrieve all relevant information with no loss of memory

- process all information, often contradictory, to form a single, integrated judgement or perception, in the same fashion, using cognitive processes that are unbiased
- convey this judgement as a final response in each survey context identically

If so, individuals with exactly the same true level of health status should then respond identically on health status surveys, across and within populations. Obviously, this optimal strategy does not exist in practice and is no surprise to those critically assessing methods used within international health and development (Zborowski 1952; Wolff & Langley 1968; Kleinman 1994; Sen 1994; Weir & Seacrest 2000). But what are the likely consequences of the various differences that can and do seep in?

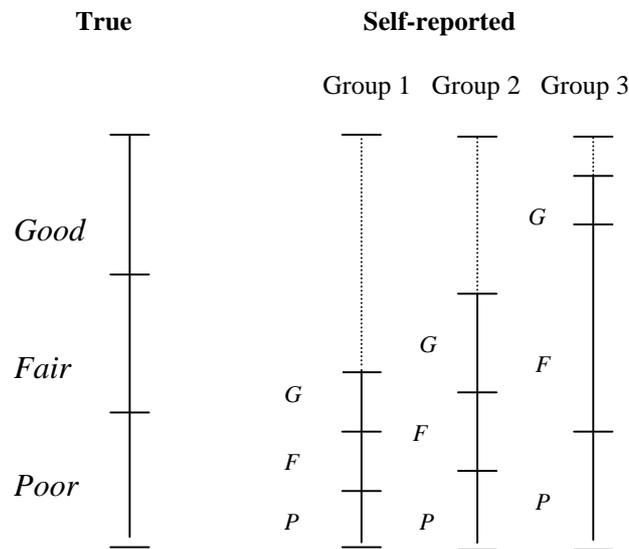
Although a complete discussion of both earlier and later work on response bias and judgmental heuristics is beyond the scope of this paper (see Kahneman, Slovic & Tversky 1982), the fact that different strategies are used to respond to questions -- and that these strategies are biased to different degrees, and are used by different individuals to different degrees -- suggests that even if questions are equivalent, data collected on identical questions assessing health status may not be equivalent and thus not comparable. These differences are not simply measurement errors. This view is in sharp contrast to those who interpret data collected on self-reported health status, across or within populations, at face value. It is also in contrast to those who focus on improving the comparability of instruments and survey methods, and not on data equivalence.

We propose to call a stricter form of validity *cross-population comparability*, or X-comparability for short. X-comparability requires measurement equivalence in terms of reliability as well as adherence to the optimal strategy for equivalence of data based on the same scale references across populations, in terms of validity. It also requires consistent reporting, discussed in the next section. We argue that estimating X-comparability requires external validation techniques that are beyond those commonly used in the development and testing of health status assessment instruments.

Differences in end-points and cut-points on scale references indicate the lack of X-comparability. For example, different experiences, expectations or norms and cognitive processes may contribute to how individuals calibrate responses. Figure 1 illustrates that both the cut-points and end-points of a given scale may differ for different groups of individuals. Compared to a scale that measures the *true* or *full range* of health status (for simplicity, represented as *good*, *fair*, and *poor* in Figure 1), the entire scale is compressed significantly for group 1, a bit less for group 2 and only slight for group 3. In other words, although individuals distinguish between *good*, *fair* and *poor*, the range of what is the potential level of health (end-points) is compressed to different degrees.

We hypothesize that individuals with low expectations for health and less exposure to what constitutes higher levels of health, may fall into Group 1. Groups 1 and 2 utilize a different range (end-points) and thus the cut-points line up differently. Nevertheless, the cut-points appear at equally appearing intervals for Group 1 and Group 2. In contrast, although the scale for Group 3 is only slightly compressed in comparison to the True scale, *fair* represents a much larger proportion of true health status than *good* or *poor*. Group 3 may be composed of individuals with higher socio-economic characteristics, who have a broader range of exposure to information on health, and higher expectations for their level of health and for the standards that distinguish *good* from *fair* health.

Figure 1. Hypothetical relation between the scaling properties of True and Self-reported Health status across demographic, socio-economic or cultural groups



If these scales are representative of the groups discussed, then individuals in group 1 may report themselves in *good* health, although their true health status may be *fair*, or even *poor*. Furthermore, an unreflective comparison of data collected from groups 1 and 3 would incorrectly assume that individuals reporting *good* from group 1 is equivalent to individuals reporting *good* in group 3. Individuals in group 3, however, are more likely to classify a broader range of health status states as *fair* in comparison to *good* and *poor*. As a result, individuals with higher socio-economic characteristics may self-report worse health in comparison to individuals with lower socio-economic characteristics.

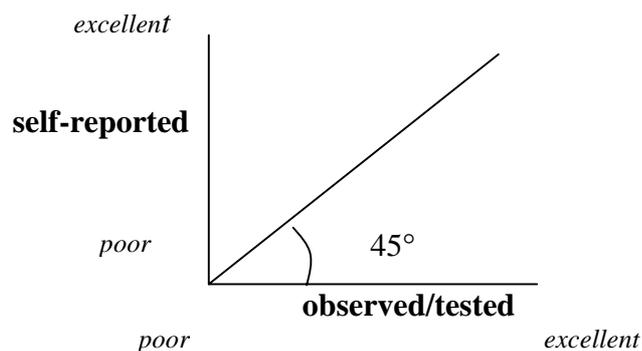
2. Differences between Self Reported, Tested and Observed

As there is no gold standard test to measure health status, the next best alternative employed is to compare observed or tested health status on a variety of domains, with what people self-report in the same or similar domains.

Health status is *consistently reported* not only if higher levels of observed or tested health are reflected in higher levels of self-reported health, but also if these are identical, i.e. on the same scale yielding the same intercept and slope (e.g., not simply a monotonic relationship). For example, if people who have poor eyesight based on vision tests report their visual ability to be worse than people who have better eyesight, then this is only a monotonic relationship. Consistent reporting is a precondition for using self-reported health status as a measure of the underlying true health status and may be considered as one requirement of X-comparability. Yet consistent reporting is not sufficient as an indicator of validity, as observed, tested and self-reported health status may be identical but still differ from true underlying health status. This is reflected by improvements in the validity of clinical assessments, tests and survey questions over time.

For example, if people who have poor eyesight based on vision tests report their visual ability to be worse than people who have better eyesight, then this is only a monotonic relationship. Figure 2 depicts the stricter criterion of consistent reporting using the same scale on both axis, in the simplified case that the relationship is linear. Throughout the range of observed health status, people in better observed health should report their health to be better, and the degree to which these reports and observations agree are perfect. In the example depicted in Figure 2, inconsistent reporting would be marked by deviations from the 45° line through the origin. Consistent reporting is a precondition for using self-reported health status as a measure of the underlying true health status and may be considered as one requirement of X-comparability. Yet consistent reporting is not sufficient as an indicator of validity, as observed, tested and self-reported health status may be identical but still differ from true underlying health status. This is reflected by improvements in the validity of clinical assessments, tests and survey questions over time.

Figure 2. Hypothetical relation between Observed and Self-Reported: Health status, if consistently reported



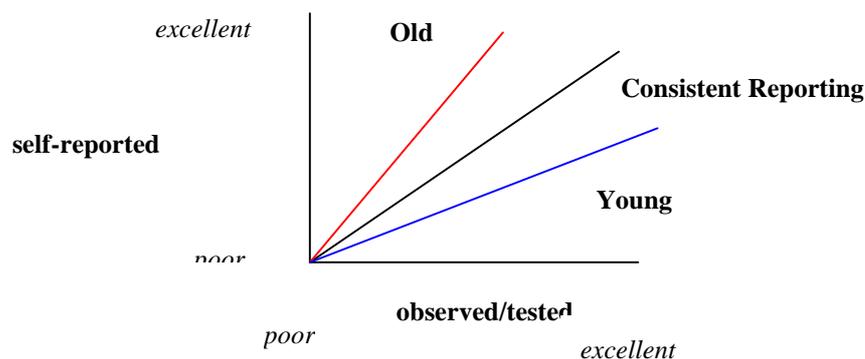
Inconsistent reporting may result from different expectations and norms for health as well as biased judgmental processes. These may differ by age, sex, and other sub-population characteristics and contribute to the gap between self-reported health status and true health status in complex ways. These differences between self-reported and observed health status are expected across:

- age groups;
- sex and gender;
- economic levels;
- education levels;
- cultural and ethnic groups;
- occupations;
- levels of health insurance and other benefits or entitlements;
- overall level of industrialization or development; or
- time periods.

A variety of factors may contribute to differences in expectations and norms for health. As discussed, some groups may have different standards for *excellent* health than other groups. Imagine that people around 20 to 25 years of age think that *excellent* mobility is being able to run a marathon while people around 75 to 80 years of age think that excellent mobility is being able to walk a kilometer without pain. If asked about the ability to engage in vigorous activities, the 22 year old who can walk only a half a kilometer before experiencing pain may report his or her health as *poor*, while the 78 year old who had the same experience might report his or her health as *good* or even *very good*.

Figure 3 shows such differences in norms and expectations based on age and the expected deviation from consistent reporting, again in the simplified case where this relationship is linear. At any level of observed health, younger people will report themselves in worse health than the oldest age groups. This is so as they perceive and report their health worse in comparison to their standard of excellence, which is higher than the standard of excellence for the oldest age groups. The gap may be smallest at poor levels of performance (where excellence is not an issue) and greatest at the highest levels. Figure 3 illustrates that both groups have the same intercept, but that the slope of the relationship between observed and self-reported health status differs from one of consistent reporting.

Figure 3. Hypothetical relation between Observed and Self-Reported Health for different age groups



Initial efforts to compare observed or measured health status with self-reported health status largely focused on estimating the validity of self-reported morbidity status in comparison to clinical or laboratory examinations or medical records (Rubin, Rosenbaum & Cobb 1956; Elinson & Trussel 1957; Krueger 1957; NCHS 1961; Woolsey, Lawrence & Balamuth 1962; Mechanic & Newton 1965; Belcher et al. 1976). These and subsequent studies in industrialized and less industrialized regions concluded that the relationship between an array of self-reported morbidity and observed morbidity is not consistent as expected (Colditz et al. 1986; Harlow & Linet 1989; Zurayk et al. 1995; Filippi et al. 1997; Mackenback, Looman & van der Meer 1996; Kaufman et al. 1999; Sadana 2000).

Several studies from different regions describe that different patterns exist among sub-populations. For example, in Argentina higher socio-economic groups reported more illness than lower socio-economic groups in household interview surveys when the reverse was true based on information external to the survey (Kroeger et al. 1988). The same pattern appears to hold true across regions. Based on data collected within the National Sample Survey of India, Round 28, residents of Kerala -- the state in India with the lowest levels of infant and child mortality and highest levels of literacy -- report the highest incidence of acute morbidity in the country.

Whereas the reverse is true for residents of Bihar -- the state often with the worst indicators concerning mortality and literacy in India -- as they report the lowest incidence of acute morbidity in the country (see Murray 1996). Several studies in Europe also document that less educated individuals under-report morbidity conditions to a greater extent than individuals with more education (Heliovaara et al. 1993; Mackenback et al. 1996). These findings are consistent with the hypothesis that greater awareness of symptoms and higher expectations for good health, increases the likelihood of reporting morbidity and health status problems. These studies provide evidence of the lack of X-comparability.

Although many studies compare the self-report of morbidity with observed clinical or laboratory findings, few studies provide evidence on the relationship between self-reported health status and observed health status measures, i.e., describing different health states beyond simply reporting morbidity or disease status. Representative of early studies in the United States, Friedsan and Martin (1961) found a positive but weak relationship between self-reported health and physician's ratings. More recently, two studies from the Netherlands and Finland found self-reported health and observed measures to be only weakly correlated (Kempen et al. 1996; Kivinen et al. 1998). These and other studies (Matthias et al. 1993; Axelsson and Helgadottir 1995) indicate that this relationship may be inconsistent and differs across sub-populations.

A complex pattern of reporting emerges from recent analysis conducted by Thomas & Frankenberg (2000), using data from the Indonesian Family Life Survey, 1993. They contrast the correlations between respondent characteristics (such as on age, sex, education levels, or per capita expenditures in the household) and self-reported health status (such as different ADL questions) with the correlations between the same characteristics and observed functioning tests (i.e., corresponding to the self-reported ADL question). The main findings are that different response patterns exist for each question and that these responses potentially reflect different normative roles and expectations. For example, for the question "carrying a heavy load," higher levels of per capita household income is associated with *more* difficulty, only for women. But that this is not the case for some other ADLs. In this example, the authors are tempted to conclude that the act of carrying a heavy load is not something higher income women view as part of their daily activities and that reporting difficulties in this area is more a reflection of the appropriateness of the activity within their peer group, rather than physical limitations.

In addition to differences based on individual characteristics, differences in the self-report of health status over time may reflect changes other than changes in health status. For example, Midthjell et al. (1992:542) find that recent efforts to estimate the validity of a population-based health interview survey eliciting diabetes diagnosis and treatment compared to medical records, yielded excellent results in comparison with a validation study 10 years earlier. The authors attribute the narrowed gap between self-reported and observed morbidity due to "better medical awareness by patients, a greater consensus on diagnostic routines, and increasing willingness of physicians to take responsibility for providing information and education for their patients about chronic diseases." Along the same lines, in an analysis of longitudinal data collected within the National Health Interview Surveys in the United States, Waidmann et al. (1995) suggest that declines in self-reported general health status among the elderly during the 1970's probably did not reflect changes in the actual health of the population, but that these declines can be interpreted as "reflecting changes in underlying social forces that influence individual awareness of attitudes toward and options for accommodating chronic health problems" (1995:280). An important dimension of health transition research investigates these changes in social and professional norms, access to health services, and individual behavior that influence the recognition and classification of symptoms and the demand for health care (Wilson & Drury 1984; Findley 1990; Caldwell and Caldwell 1991; Johansson 1991, 1992; Riley 1992; Annas 1993; Murray 1996).

The biases associated with the self-report of health status and the inconsistent patterns of reporting are a complex function of expectations, norms, exposure to health services, information and judgmental strategies. We do not suggest that these biases render individual assessments of health status less important, but that these biases reduce the cross-population comparability of data.

3. Using observed data to improve health status measurement

Several strategies are currently being pursued in order to improve the comparability of questions and to a significantly lesser degree, the cross-population comparability of data collected within household interview surveys. WHO is currently investigating ways of using observed data (performance tests) to calibrate self-report health status data and improve the cross-population comparability of such data.

Although IRT approaches are informative, the response patterns may still include reporting biases that serve to reduce X-comparability, as the self-report on other questions in a scale are used to anchor and compare the self-report on each question within a scale, i.e. there is no external approach to scale calibration as the underlying latent construct is measured by self-reported items. For example, if reporting biases are similar in two populations (i.e., similar inconsistent patterns), IRT or other approaches to test for homogeneous item functioning will show that item parameters are similar, and thus assume measurement equivalence despite reporting biases. Those using IRT do not address this issue while assessing the equivalence of item calibrations across different contexts. This is apparent in the remedy suggested when there is evidence that items behave differently in different populations: a call to improve the translation of questions is evoked as the means to reach homogeneous item functioning (see Razcek et al. 1998 or Bjorner et al. 1998), rather than adjusting for potential differences in cut-points and end-points or inconsistent reporting patterns.

WHO plans a comprehensive strategy to enhance the comparability of data collected on health status from household interview surveys, as a part of the operational approach of the WHO Common Framework for Measuring and Reporting on the Health of Populations. The first component of this strategy is to intensify our data search to uncover existing nationally representative surveys from countries that have not been previously reviewed for the comparability of data. The second component is to identify and test innovative methodologies for data collection and analysis that may estimate X-comparability and offer approaches to adjust data to overcome at least some of the limitations of self-reported data discussed in this paper. The third component is to facilitate comparability studies and support the primary collection of data particularly in regions or sub-regions where few nationally representative data sets are found. Our overall objective of this strategy is to move towards cross-population comparability. We briefly outline the second component strategy relating to the use of observed data.

Consistent reporting. As there is no gold standard test to measure the true level of health status, we propose that consistent reporting across countries is documented by comparing self-reported health status and tested health status. Given that professional ratings or observations show similar biases as self-reported assessments, we prefer to utilize performance tests across countries as a means to evaluate consistent reporting. Ideally we would recommend to do these comparisons for each candidate domain. However, we are faced with two practical constraints. The first is that performance tests that are independent of other factors, such as culture or education levels, do not necessarily exist for all domains, and for other domains performance tests are inappropriate. The second is that within the time and resource constraints of nationally

representative surveys, it is necessary to select a few domains as an initial step to gauge differences in consistent reporting across countries and limitations in the comparison of data collected.

We propose to do so for the domains of vision, cognition and mobility. Standardized performance tests for each of these domains exist that are relatively objective and have been used in cross-national settings. These include Snellen vision tests; face recognition, shape cancellation and word recall tests; and composite gross mobility tests. If patterns deviating from consistent reporting are noted for different sub-groups, such as by age, sex, or socio-economic class, or across populations, evidence will be gained to justify calibrating self-reported health status across sub-populations or populations, in order to improve data equivalence. Tests for consistent reporting are an extension of current criterion based approaches to estimate the validity of health status assessment approaches.

End-points and cut-points. Various factors contribute to differences in end-points and cut-points on scales. In order to gauge how much variation exists across populations concerning differences in scale references, we suggest building in calibration techniques to establish the equivalence of cut-points, if not also end-points. Both internal and external means to do so should be integrated within surveys.

Irrespective of the scale or the magnitude of differences between cut-points established by external criteria, we will then document if the cut-points people use to judge domains of health status described in the series of vignettes differ across countries. Some populations may place more vignettes in one category or another, reflecting differences in cut-points. By combining information based on external criteria in setting meaningful end-points and cut-points, with how different levels of health are categorized, we may then calibrate responses across scales and enhance the X-comparability of data. This focus on enhancing X-comparability requires external calibrations techniques, and this may be considered as a new dimension in cross-population research in this area.

Support comparability studies and primary data collection. As a step forward in this direction, WHO, in conjunction with collaborators, has drafted a set of generic domains and survey questions for further testing, and developed a survey design that will test and refine several of the strategies discussed to enhance X-comparability. A series of comparability studies in selected countries in each of the six WHO regions are being initiated. The first phase of studies will be directly supported by WHO: pilot testing is currently underway in 10 countries¹ in preparation for nationally representative sample surveys of the non-institutionalized population from urban and rural areas in each country and an additional 60 countries through postal surveys.

A second phase of comparability studies will be conducted in a broader range of countries, with WHO providing technical support along with standardized methodologies. It is intended that the main result from these comparability studies will be the development of a standardized module on health status that incorporates strategies to estimate and adjust for X-comparability, in addition to meeting other criteria for validity and reliability.

Acknowledgments

Figures in section 2 are based on previous collaborative work with David Cutler, Department of Economics, Harvard University. This paper for presentation in Ottawa is extracted from a larger paper (Sadana et al. 2000), available on the following website: www.who.int/evidence select discussion papers, #15.

¹ These include: China, Columbia, Egypt, Georgia, Nigeria, India, Indonesia, Lebanon, Slovakia and Turkey

References

- Anderson RT, Aaronson NT and Wilkin D 1993. Critical review of the international assessments of health-related quality of life. *Quality of Life Research* 2:369-395.
- Annas J (1993). Women and quality of life: two norms or one? In *The quality of life* Nussbaum M and Sen A (eds.) Clarendon Press, Oxford, pp.279-296.
- Axelsson G, Helgadóttir S (1995). Comparison of oral health data from self-administered questionnaire and clinical examination. *Com Dentistry and Oral Epidemiology*, 23:365-8
- Belcher DW, Newmann AK, Wurapa FK, Lourie IM (1976). Comparison of morbidity interviews with a health examination survey in rural Africa. *American Journal of Tropical Medicine and Hygiene* 23: 751-758.
- Bergner M, Rothman (1987). Health status measures: an overview and guide for selection. *Ann Rev Public Health* 8:191-210.
- Bjorner JB, Kreiner S, Ware JE. et al. (1998). Differential item functioning in the Danish translation of the SF-36. *J Clin Epidemiology* 51(11):1189-1202
- Caldwell J, Caldwell P (1991). What have we learnt about the cultural, social and behavioral determinants of health? From selected readings to the first health transition workshop. *Health Transition Review* 1:3-20.
- Colditz GA, Stampfer MJ, Willet WC et al. (1986) Validation of questionnaire information on risk factors and disease outcomes in a prospective cohort study of women. *Amer J Epi* 123(5):894-900.
- Cooke DJ and Michie C 1999. Psychopathy across cultures: North America and Scotland compared. *Journal of Abnormal Psychology*. 108(1):58-68
- Eurostat 1997. Self-reported health in the European Community. *Statistics in Focus, Population and social conditions*. ISSN 1024-4352
- Filippi V, Marshall T, Bulut A. et al. (1997). Asking questions about women's reproductive health: validity and reliability of survey findings from Istanbul. *Topical Medicine and International Health*, 2(1): 47-56
- Findley SE (1990). Social reflections of changing morbidity during health transitions. *Proceedings from the Health Transition Workshop*, Vol 1., Australia.
- Frankenberg E, Thomas D, Beegle K (1999). *The Real Costs of Indonesia's Economic Crisis: Preliminary Findings from the Indonesia Family Life Surveys*. Labor and Population Program Working Paper Series 99-04, The RAND Corporation, Santa Monica, California.
- Friedsam HJ, Martin HW (1961). A comparison of self and physicians' health ratings in an older population. *J Health & Human Behavior*, 179-183.
- Groot W 2000. Adaptation and scale of reference bias in self-assessments of quality of life. *J Health Economics* 19:403-420.
- Hambleton RK and Swaminathan H 1985. *Item response theory: principles and applications*. Kulwer Nijhoff, Boston.
- Hambleton RK, Swaminathan H and Rogers HJ 1991. *Fundamentals of item response theory*. Sage, Newbury Park, CA.

- Harlow DE, Linet MS (1989) Agreement between questionnaire data and medical records. *Amer J Epi* 129(2):233-248.
- Heliövaara M, Aromaa A, Klaukka T. et al. (1993). Reliability and validity of interview data on chronic diseases. *J Clin Epidemiology* 46(2):181-191.
- Idler E.L., Hudson S.V., and Leventhal H. 1999 The Meanings of Self-Ratings of Health: A Qualitative and Quantitative Approach. *Research in Ageing*
- Jobe JB, Mingay DJ (1989). Cognitive Research Improves Questionnaires. *Am J Pub Health* 79(8):1053-1055
- Johansson SR 1991. The health transition: the cultural inflation of morbidity during the decline of mortality. *Health transition review* 1(1):39-68.
- Kahneman D, Slovic P, Tversky A (eds). 1982. *Judgment under uncertainty: heuristics and biases*, Cambridge University Press, Cambridge
- Kaufman J. et al. A study of field-based methods for diagnosing reproductive tract infections in Rural Yunnan Province, China. *Studies in Family Planning*, 1999, 39(2):112–119
- Keller SD, Ware JE, Bentler PM et al. 1998. Use of Structural Equation Modeling to Test the Construct Validity of the SF-36 Health Survey in Ten Countries: Results from the IQOLA Project. *J Clin Epi*, 51(11):1179-1188
- Kempen GI, van Heuvelen MJ, van den Brink RH et al. 1996. Factors affecting contrasting results between self-reported and performance-based levels of physical limitations. *Age & Ageing* 25(6):458-64
- Kleinman A, Eisenberg L and Good B 1978. Culture, illness and care: clinical lessons from anthropologic and cross-cultural research. *Annals of Internal Medicine* 88: 251-258
- Kleinman A 1994. An anthropological perspective on objectivity: observation, categorization, and the assessment of suffering. In *Health and social change in international perspective*, LC Chen, A Kleinman, NC Ware (eds.). Harvard series on population and international health, Harvard School of Public Health, Boston, pp. 129-138.
- Kivinen, P, Halonen P, Eronen M, Nissinen A 1998. Self-rated health, physician-rated health and associated factors among elderly men: the Finnish cohorts of the Seven Countries Study, *Age & Ageing*, 27:41-47.
- Kroeger A, Zurita A, Perez-Samaniego C, Berg H, 1988. Illness perception and use of health services in North-East Argentina. *Health Policy and Planning* 3: 141-151.
- Krosnick JA 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5:213-236
- Krueger DE (1957). Measurement of prevalence of chronic disease by household interviews and clinical evaluations. *AJPH* 47: 953-960.
- Mackenbach J.P., Looman C.W.N. van der Meer J.B.W. 1996. Differences in the Misreporting of Chronic Conditions, by Level of Education: The Effect on Inequalities in Prevalence Rates. *American Journal of Public Health* 86(5):706-711.
- Matthias RE, Atchison KA, Schweitzer SO, Lubben JE, Mayer-Oakes A, De Jong F (1993). Comparisons between dentist ratings and self-ratings of dental appearance in an elderly population. *Special care in Dentistry*, 13(2):53-60.

- Mechanic D, Newton M (1965). Some problems in the analysis of morbidity data. *J*
- Murray CJL and Chen LC 1992. Understanding Morbidity Change. *Population and Development Review*, 18(3):481-503
- Murray CJL 1996. Epidemiology and morbidity transitions in India. In *Health, Poverty and Development in India*, eds. DasGupta M, Chen LC and Krishnan TN, Oxford University Press, Delhi, 122-147.
- Murray CJL, Salomon J, Mathers CD, Lopez A, Lozano R (2000). Summary Measures of Population Health. Geneva, World Health Organization (in preparation).
- NCHS 1961. Health interview responses compared with medical records. Vital Health Statistics, Series D, No. 5. Washington, D.C. Public Health Service publication, United States National Centre for Health Statistics.
- Nunnally JC and Bernstein IR 1994. *Psychometric Theory*. Third edition. McGraw Hill, New York.
- Raczek AE, Ware JE, Bjorner JB, Gandek B, et al. 1998. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA project. *J Clin Epidemiology* 51(11):1203-1214
- Rasmussen N, Gudex C and Christensen S 1999. *Survey Data on Disability*, Eurostat Working Papers, Population and Social Conditions 3/1999/E/n 20, European Commission, Luxembourg.
- Riley JC (1992). From a high mortality regime to a high morbidity regime: is culture everything in sickness? *Health transition review* 2(1):71-77.
- Robine JM, Jagger C and Egidi V 2000. *Selection of a Coherent Set of Health Indicators, Final draft A First Step Towards A User's Guide to Health Expectancies for the European Union* Montpellier (France), Euro-REVES
- Rubin T, Rosenbaum J, Cobb S (1956). The use of interview data for the detection of associations in field studies. *J Chron Dis* 4:253-267
- Sadana R (2000). Measuring Reproductive Health: Review of community-based approaches to assess morbidity. *The Bulletin of the World Health Organization* 78(5):640-654
- Sen A 1994. Objectivity and position: assessment of health and well-being. In *Health and social change in international perspective*, LC Chen, A Kleinman, NC Ware (eds.). Harvard series on population and international health, Harvard School of Public Health, Boston, pp. 115-128.
- Streiner DL and Norman GR 1995. *Health Measurement Scales: a practical guide to their development and use*. Second edition. Oxford University Press.
- Thomas D and Frankenberg E. 2000. The measurement and interpretation of health in social surveys. In CJL Murray, CD Mathers, AD Lopez, J Salomon, R Lozano (eds). *Summary measures of population health*, Geneva, World Health Organization (under preparation).
- Tourangeau R 1984. Cognitive sciences and survey methods. In *Cognitive aspects of survey methodology: building a bridge between disciplines*. T Jabine, M Straf, J Tanur and R Tourangeau (eds). Pp. 73-100. National Academy Press, Washington D.C.
- Tversky A, Kahneman D (1974). Judgment under uncertainty: heuristics and biases. *Science* 185:1124-1131

- Waidmann T, Bound J, Schoenbaum M (1995). The Illusion of Failure: Trends in the Self-report of health of the US Elderly. *The Milbank Quarterly* 73(2): 253-287.
- Ware JE, Keller SD, Gandek B, Brazier JE, Sullivan M et al. 1995. Evaluating translations of health status questionnaires: methods for the IQOLA project. *Int J Technol Assess Health Care* 11(3):525-551
- Ware JE, Keller SD (1996). Interpreting general health measures. In *Quality of life and pharmoeconomics in Clinical Trails*, second edition, Spilker B (ed). Lippencott-Raven Publishers, Philadelphia.
- Ware JE, Gandek BL, Keller SD, the IQOLA Project Group (1996). Evaluating instruments used cross-nationally: methods from the IQOLA project. In *Quality of life and pharmoeconomics in Clinical Trails*, second edition, Spilker B (ed). Lippencott-Raven Publishers, Philadelphia.
- Weir C and Seacrest M 2000. Developmental differences in understanding of balance scales in the United States and Zimbabwe. *J Genetic Psychology* 161(1):5-21
- WHOQOL Group 1998. The World Health Organization Quality of Life Assessment (WHOQOL): development and general psychometric properties. *Social Science & Medicine*. 46(12):1569-85
- Wilson RW, Drury TF (1984). Interpreting trends in illness and disability: health statistics and health status. *Ann. Review public Health* 5: 83-196.
- Wolff BB, Langley S (1968). Cultural factors and the response to pain. A review. *American Anthropologist* 79: 494-501.
- Woolsey TD, Lawrence PS, Balamuth E. An evaluation of chronic disease prevalence data from the health interview survey. *American Journal of Public Health*, 1962, 52(10):1631-1637
- Zborowski M (1952). Cultural components in response to pain. *Journal of Social Issues* 8: 16-30.
- Zurayk H, Khattab H, Younis N et al. 1995. Comparing women's reports with medical diagnosis of reproductive morbidity conditions in rural Egypt. *Studies in Family Planning* 26(1):14-21.